Predicting the Survival Of The Titanic Passengers

Asif Mohammad   
Dept. Of Computer Science &Engineering  
East West UniversityAftabnagar, Dhaka  
asifratul45@gmail.com

Shifat Bin Habib  
Dept. Of Computer Science &Engineering  
East West UniversityAftabnagar, Dhaka  
shifat514@gmail.com Mahruf Zaman Utso   
Dept. Of Computer Science &Engineering  
East West UniversityAftabnagar, Dhaka  
mahruf.z44@gmail.com

Nusrat Sharmin Anika   
Dept. Of Computer Science &Engineering  
East West UniversityAftabnagar, Dhaka  
nusrat.anika15@gmail.com

***Abstract***— **It has been over a hundred years since the Titanic disaster happened; still, it has created interest among data scientists and researchers about acknowledging the idea that why some passengers survived the shipwreck while the demise of the others. A big factor for the loss of lives on the Titanic is that the quantity of lifeboats available on the ship was not enough for everyone. Although people of all classes, genders & ages had lost their lives on a fateful day from some interesting observations, it could be shown that among the survivor’s women and children were getting the priority to be rescued first. In our research with our selected dataset, we are attempting to make a prediction on the survival of the titanic passengers. We are applying linear regression, logistic regression, SVM, decision tree, and naïve Bayes algorithms so that we can predict the possible outcome.**

**Keywords: Machine Learning, Titanic, Prediction, Regression, Decision tree, Naïve Bayes.**

Introduction

The Titanic disaster is considered the most famous shipwrecks in world history which occurred on April 15th, 1912. On a fateful day, during its maiden voyage from Southampton to New York City, it struck an iceberg and sank in the North Atlantic Ocean. [1] There were approximately 2224 passengers and crew aboard the ship. More than 1500 people lost their lives when the Titanic sank in the Ocean.  Unfortunately, there were not enough lifeboats for everyone on board and this caused the death of the passengers and crew. [2] Titanic was considered as an “unsinkable ship” but due to poor management and misjudgment, it caused the shipwreck.  Although many people lost their lives, some survived which is the reason for arising various speculations regarding the surviving passengers in the Titanic disaster.

Our aim is to predict the survival of Titanic passengers using machine learning. The main purpose of machine learning is to provide the system the ability to learn automatically and also improve its features from experience without being bluntly programmed. It mainly focuses on the development of computer programs by accessing data and using it to learn for themselves. [3]

**Related Work**

For predicting the survival of the titanic passengers we have used a machine learning algorithm on the dataset. Titanic is used as one of the most popular datasets for data science. [4] In this dataset, there are various records including the passenger’s information and also the survived passenger information.  Machine learning divides the data into two parts:  one is for training another one for testing the data. The given attributes given in the dataset (such as passenger, name, age, sex, etc.) are used as features. For analyzing the data, the effects of the features need to be explored and depending on the necessities features can be removed.

Over the years, many researchers have been studying and analyzing the titanic dataset, using different machine learning algorithms for example KNN, Regression, Random forest, Classification, and many more. Also, different languages and tools have been used to implement these algorithms. From a study conducted by Farag, using the Decision tree and Gaussian Naïve Bayes algorithm where we see that The Decision Tree algorithm has accurately 90.01% of the survival of passengers, while the Gaussian Naïve Bayes calculate 92.52% accuracy in prediction for the survival of passengers [5]. Our approach is centered on SQL for executing algorithms, as for tools we are using oracle Machine Learning. The algorithms used for predicting the survival of the passengers are Linear regression, Logistic regression, SVM, Decision tree, and Naïve Bayes.

**Methodology**

We chose to implement Linear Regression, Logistic Regression, Support Vector Machines, Naive Bayes, and Decision Tree algorithms on our dataset to predict the survivors.

1. Linear Regression

In machine learning Linear Regression is a supervised machine learning algorithm that is used for Classification and Regression Analysis. Here the output is predicted to be continuous and it has a constant slope. Linear Regression predicts the values within a continuous range. There are two types of linear regression. One is Simple Linear Regression and the other is Multivariate Linear Regression. We used Simple Linear Regression algorithm to implement in our Dataset

Simple Linear Regression is a model which indicates the relationship between the dependent and independent variable it fits a linear equation to the selected data. The equation can be described as

Where X is the independent or explanatory variable represents the input data, Y is the dependent variable represents the prediction, m is the slope of this line and c is the y intercept. The algorithm will learn the value of m and c to generate the most accurate predictions

B. Logistic Regression

Logistic Regression is a kind of regression that is used for predicting a dependent variable where a set of independent variables are given and also the dependent variable is categorical. Its underline technique is quite similar to linear regression. In logistic regression, the categorical dependent variable can hold values like 1 or 0, yes or no, and so on. We cannot expect categorical results from linear regression because we input continuous variables into it. Logistic regression predicts the probability of the outcome in binary 0 and 1.

 Here, y is the probability of an event happening which we are trying to predict. X1 and x2 are the independent variables and it determines the occurrence of y event. c is the constant term that indicates the probability of the event happening when no other factors are considered.

C. Support Vector Machine

Support Vector Machine is a supervised machine learning technique that is used for classification and regression analysis. It was established by Vapnik in 1995 [6]. Support Vector Machine algorithm separates different classes of training data by an optimal hyperplane or decision boundary [7]. The hyperplane separates the different classes of training data by a maximal margin. If the training data are non-linearly distributed, then SVM (Support Vector Machine) maps those Low Dimensional Space Data to High Dimensional Space data by using kernel function and separates those different classes of data by hyperplane considering the maximal margin.

D. Naive Bayes

Naive Bayes achieves efficient and fast classification in machine learning applications. Naïve Bayes performs well on high dimensional and complex datasets. Naive Bayes algorithm was developed based on the Bayes theorem.

Consider that an event E corresponds to (Feature = value) and another event D corresponds to (Class = c). The goal is to determine probability P (D|E) which means determining the probability of D given that event E.

Bayes theorem states the following:

If A and B events are independent of one another and conditional on a third event F, then it follows as below:

E.    Decision Tree

Decision tree is known for using in classification and regression. Decision trees are modeled with a set of hierarchical decisions on dataset attributes. The decision at each node of the tree is referred to as the split criterion. This decision is a condition on single or multiple feature attributes in the training dataset. The training dataset is divided into multiple parts based on the split criterion.

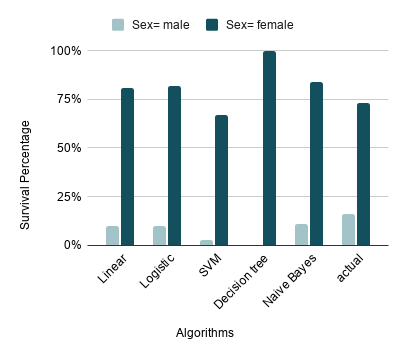
Each node represents a subset of the dataset defined by the combination of split criteria in its upper nodes. Decision tree is mainly a sequence of splits in a top-down fashion to create nodes at the leaf level. These split criteria actually determine the output of test data results. Based on these split criteria in each node decision tree predicts the outcome of a query.

|  |  |
| --- | --- |
| ATTRIBUTE\_NAME | ATTRIBUTE\_TYPE |
| AGE | NUMERICAL |
| CABIN | CATEGORICAL |
| EMBARKED | CATEGORICAL |
| FARE | NUMERICAL |
| NAME | CATEGORICAL |
| PARCH | NUMERICAL |
| PCLASS | NUMERICAL |
| SEX | CATEGORICAL |
| SIBSP | NUMERICAL |
| SURVIVED | NUMERICAL |
| TICKET | CATEGORICAL |

Table 1: Description of attribute feature.

We have worked with the famous titanic dataset [8]. There are a total of 12 attributes. The attributes of this dataset are passenger, survived, Pclass Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. Descriptions of the attributes is shown in the table 1.

We divided the dataset into train dataset which is 70% of the whole dataset and test dataset which is the rest of the 30%.  We have chosen some algorithms to predict the survival of the passengers.   We have applied logistic regression, linear regression, decision tree, Naive Bayes and support vector machines,.



After applying the algorithms to the whole dataset, we have worked with some attributes. These attributes are Sex, Age, Pclass. We divided the Age into 4 categories. These categories are 0-15, 16-30, 31-60, and lastly 61 and above. We added SibSp and Parch attributes together and made a feature named Family-size We divided Family-size into two parts. One part is lower than 5 people and another one that is greater than or equal to 5 people.

Then for each of these four attributes and features, we have compared the predicted survival rate and the actual survival rate for each algorithm we have applied.

The algorithms we used were implemented in the Oracle machine Learning notebook and we used PL SQL programming language for our project.

After that, we compared the accuracy of all the algorithms.

**Data Findings & Result**

We analyzed the following features:

Sex:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| features | Linear | Logistic | SVM | Decision tree | Naive Bayes | actual |
| Sex= male | 10% | 10% | 3% | 0% | 11% | 16% |
| Sex=  female | 81% | 82% | 67% | 100% | 84% | 73% |

Table 2: Survival Table of five proposed algorithms and actual data in terms of Sex

There are two values for Sex attributes one is male and another is female. In Table 2, we can see the prediction of the survival rate of passengers according to the 5 proposed algorithms (Linear Regression, Logistic Regression, Support Vector Machines, Naive Bayes, and decision tree) with the respect to sex attribute values. We have also calculated the actual survival rate for both male and female from the whole dataset which is also shown in table 2. If we compare the survival rate of 5 algorithms with the actual survival rate then we can see that Naive Bayes, Logistic, and Linear Regression algorithms have the closest survival rate as the actual survival rate for the male value of Sex attribute. For the Female value of the Sex attribute, all the algorithm’s survival rate is as close as the actual rate except for the Decision tree algorithm which survival rate seems to differ from the actual.

If we show the above facts in a bar chart, then the figure shall like figure 1:

Figure 1: Survival Bar chart of five proposed algorithms and actual data in terms of class sex

After predicting the survivors from all the algorithms, we also found that the sex=female has survived more than sex=male and the actual dataset also shows the same result.

Age:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| features | Linear  Regression | Logistic  Regression | SVM | Decision tree | Naive Bayes | actual |
| Age  (0-15) | 55% | 59% | 55% | 59% | 64% | 55% |
| Age  (16-30) | 44% | 44% | 32% | 42% | 39% | 38% |
| Age  (31-60) | 29% | 30% | 25% | 33% | 38% | 38% |
| Age  (61-80) | 0 | 0 | 13 | 13% | 38% | 25% |

Table 3: Survival Table of five proposed algorithms and actual data in terms of Age

Here the age limit in the dataset was from 0 to 80. We divided the data of age between 4 groups which is Age (0-15), Age (15-30), Age (31-60) and Age (60-80) to determine the survival rate. In table 3, we can see the prediction of the survival rate of passengers according to the 5 proposed algorithms (Linear Regression, Logistic Regression, Support Vector Machines, Naive Bayes, and decision tree) with the respect to age attribute groups. We have also calculated the actual survival rate for those attribute age groups from the whole dataset which are also shown in table 2. If we compare the survival rate of 5 algorithms with the actual survival rate, then we can see that for age group (0-15) and (15-30) all proposed 5 algorithms have the closet survival rate as the actual survival rate. For the age group (16-30) all the algorithm’s survival rates are as close as the actual rate except for the Support Vector Machine algorithm which the survival rate seems to differ from the actual. For the age group (61-80) None of the proposed algorithms seems to be as close to the survival rate as the actual survival rate. From the dataset,

If we show the above facts in a bar chart, then the figure shall like figure 2:

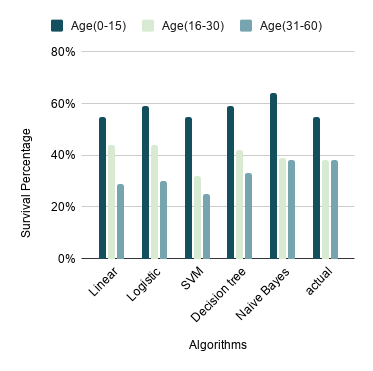


Figure 2: Survival Bar chart of five proposed algorithms and actual data in terms of class Age.

After predicting the survivors from all the algorithms, we also found that the age group of (0-15) has survived more than other age groups and the actual dataset also shows the same result.

Class:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| features | Linear | Logistic | vector | Decision tree | Naive Bayes | actual |
| class-1 | 57% | 62% | 49% | 43% | 67% | 59% |
| class-2 | 51% | 49% | 47% | 47% | 51% | 53% |
| class-3 | 23% | 22% | 9% | 32% | 21% | 22% |

Table 4: Survival Table of five proposed algorithms and actual data in terms of class.

There are three values for class attributes which are 1, 2, and 3. In table 3, we can see the prediction of the survival rate of passengers according to the 5 proposed algorithms (Linear Regression, Logistic Regression, Support Vector Machines, Naive Bayes, and decision tree) with the respect to class attribute values. We have also calculated the actual survival rate for those class attribute values from the whole dataset which are also showed in table 4. If we compare the survival rate of 5 algorithms with the actual survival rate, then we can see that for class=1 attribute values all the algorithm’s survival rate is as close as the actual rate except for Support Vector Machine and Decision tree algorithm which survival rate seems to be different from the actual. For class=2 attribute values, all proposed 5 algorithms have the closest survival rate as the actual survival rate. For class=3 attribute values all the algorithm’s survival rates are as close as the actual rate except for the Support Vector Machine and Decision tree algorithm which the survival rate seems to be different from the actual.

If we show the above facts in a bar chart, then the figure shall like figure 3:

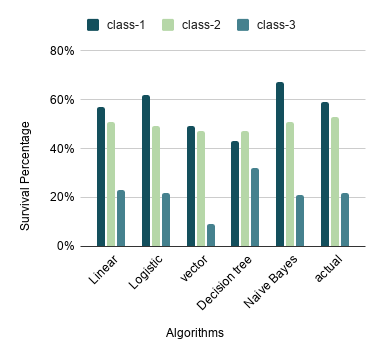


Figure 3: Survival Bar chart of five proposed algorithms and actual data in terms of class.

After predicting the survivors from all the algorithms, we also found that the class-1 has survived more than class-1 and class-2 and the actual dataset also shows the same result.

Family-size:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| features | Linear | Logistic | vector | Decision tree | Naive Bayes | actual |
| Family-size< 3 | 37% | 38% | 24% | 33% | 33% | 38% |
| Family-size>=3 | 33% | 33% | 45% | 67% | 73% | 33% |

Table 5: Survival Table of five proposed algorithms and actual data in terms of Family size

“Family-size” is created in After adding to the value of the Sibsp feature and Parch feature, we get the Family-size feature. After that, this feature is distinguished into two groups. The first group contains the passenger which family-size is less than 3 and the second group contains the passenger which family-size is greater than 3. In table 4, we can see the prediction of the survival rate of passengers according to the 5 proposed algorithms (Linear Regression, Logistic Regression, Support Vector Machines, Naive Bayes, and decision tree) with the respect to family-size attribute values. We have also calculated the actual survival rate for those family-size attribute values from the whole dataset which are also showed in table 5. If we compare the survival rate of 5 algorithms with the actual survival rate, then we can see that for Family-size< 3 attribute values all the algorithm’s survival rate is as close as the actual rate except for Support Vector Machine algorithm which survival rate seems to be different from the actual. Logistic and Linear Regression algorithms have the closest survival rate as the actual survival rate Family-size>= 3 attribute values.

If we show the above facts in a bar chart, then the figure shall like figure 4:

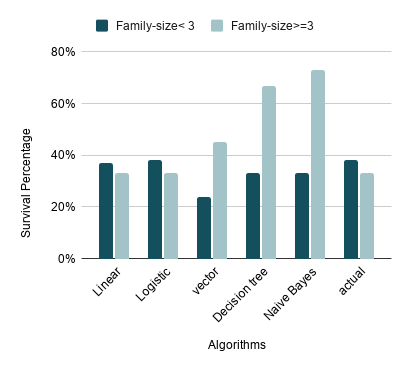


Figure 4: Survival Bar chart of five proposed algorithms and actual data in terms of family-size.

After predicting the survivors from all the algorithms, we also found that the survival rate of both family-size<3, group and family-size>=3 group has same survival rate for Linear Regression, Logistic, Regression and the actual dataset but SVM, Decision Tree and Naive Bayes predicts that family-size>=3 survived more than family-size <3

If we show the overall accuracy of 5 proposed algorithms in a bar chart, then figure should look like figure 5.

Figure 5: Overall Accuracy Graph Bar Chart

If we look at the following figure 5, we can see that Linear Regression has 85% accuracy overall which is the highest, Linear Regression has 84% accuracy, Support Vector Machine has 82% accuracy, Naive Bayes has 80% accuracy and Decision Tree has 85% accuracy.

**Discussion**

From the above calculations and results we can interpret that among the five algorithms used for predicting the surviving passengers, regression has provided higher accuracy for the given dataset. Linear Regression has the highest accuracy at 85 percentage of the dataset.

Although for the selected algorithms we have predicted the outcomes accurately but still there are some limitations in our project. As we can see for Linear Regression we get 85% accuracy, Logistic Regression got 84%, Support Vector Machine got 82%, Decision Tree got 80% and for Naïve Bayes we got 78% accuracy these accuracies the high according to some standards but there is high possibility that if we could use any other algorithms then we might get higher accuracy prediction than the existing ones.

Our project involves the implementation of data analysis and machine learning. In future we might add more algorithms for predicting, algorithms which might give us higher accuracy than existing ones. We can use various other machine learning techniques such as KNN, Adaboost, Random Forest and many others [9].

**Conclusion**

In the era of the knowledge-based world, machine learning algorithms are important to obtain valuable results from the raw and missing data. In our research paper, we proposed 5 different machine learning algorithms to predict the number of survivors from the titanic disaster. For each implemented method or algorithm we calculated the accuracy of the prediction with respect to the actual result. While calculating the accuracy, we also notice that for different features we get different accuracy. So, from this we can say that the features which are chosen for calculation or predicting accuracy can also make a great difference in predicting the survival of passengers of the titanic.

For a knowledge based world, it is very important to obtain valuable results from the raw data which we obtain from the given dataset. Finally, by using the dataset on machine learning method we can obtain the final outcome.

##### **References**

1. [RMS Titanic – Wikipedia, The Free Encyclopedia](https://en.wikipedia.org/wiki/Titanic)
2. <https://www.kaggle.com/c/titanic>
3. [What is Machine Learning? A definition](https://www.expert.ai/blog/machine-learning-definition/)
4. Ekinci, E. Omurca, and N. Acun. "A comparative study on machine learning techniques using Titanic dataset." *7th international conference on advanced technologies*. 2018.
5. <https://www.semanticscholar.org/paper/Predicting-the-Survivors-of-the-Titanic-Kaggle%2C-Farag-Hassan/2bf55df51548d6260a26f8ac45b370bc573c3c9f#citing-papers>
6. <https://en.wikipedia.org/wiki/Support-vector_machine>
7. <https://link.springer.com/article/10.1007/BF00994018>
8. Titanic dataset Link: <https://gist.github.com/michhar/2dfd2de0d4f8727f873422c5d959fff5?fbclid=IwAR3x7wo-eG_6XYmEDEHCQk1PvDLQQyUZrLxt2CqI-MRhr0n97JCJCIqINHE>
9. Singh, Aakriti, Shipra Saraswat, and Neetu Faujdar. "Analyzing Titanic disaster using machine learning algorithms." 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published**